Coping with Problems of Unicoded Traditional Mongolian

Boli Wang[1], Xiaodong Shi[1,2,3] (✉), Yidong Chen[1]

[1] Department of Cognitive Science, Xiamen University, Xiamen, China
[2] Collaborative Innovation Center for Peaceful Development of Cross-Strait Relations, Xiamen University, Xiamen, China
[3] Fujian Province Key Laboratory for Brain-inspired Computing, Xiamen University, Xiamen, China
mandel@xmu.edu.cn

Abstract. Traditional Mongolian Unicode Encoding has serious problems as several pairs of vowels with the same glyphs but different pronunciations are coded differently. We expose the severity of the problem by examples from our Mongolian corpus and propose two ways to alleviate the problem: first, developing a publicly available Mongolian input method that can help users to choose the correct encoding and second, a normalization method to solve the data sparseness problems caused by the proliferation of homographs. Experiments in search engines and statistical machine translation show that our methods are effective.

Keywords: Traditional Mongolian Script, Homographs, Input Method, Normalization

1 Introduction

Although traditional Mongolian script was standardized in ISO/IEC 10646 and Unicode in 1999, Unicoded traditional Mongolian is still not the prevalent encoding among the various forms of Mongolian encodings used in China[1], where traditional Mongolian script is the national standard, as opposed to the Cyrillic alphabet used in Outer Mongolia. This situation is due to several reasons, one of which is that there was no support for Unicoded traditional Mongolian from the major OS vendors until the release of Windows Vista in 2007. However, the way traditional Mongolian is standardized in Unicode presents many obstacles to its effective use and handling by the computer and the paper is an attempt to remedy the problems caused by the Unicode encoding of traditional Mongolian.

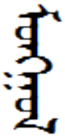
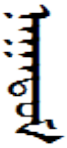
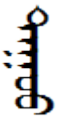
As far as we know, characters with the same glyph appearance have the same internal Unicode encoding in all languages except traditional Mongolian script, where the vowels

[1] The authors can scarcely find a Unicoded traditional Mongolian web site in the year 2014, although things began to change starting from the year 2015.

adfa, p. 1, 2011.

o and *e* have the same glyph[2] (see [2]) but different encodings (U+1823, U+1824), and *oe* and *ue* also have the same glyph[3] but are coded as (U+1825, U+1826) [3], cf. the Chinese character 行 which has different pronunciations (xíng, háng) and meanings but has the unique Unicode code point (U+884C). Another problem is that the consonant *ang* (U+1829, normally transliterated as *ng*[4]) is redundantly encoded as its glyph is often the same as the consonants *na* (U+1828) and *ga* (U+182D) joined together (thus also transliterated as *ng*). This creates lots of problems, as many words with identical glyphs but different Unicode encodings, henceforth called *homographs*, are generated by the computer users using various input methods. Table 1 lists some of the homographs we find in our publicly available Unicode Mongolian web corpus[5] with 150 million words:

Table 1. Some common Mongolian words with exact glyphs but different encodings in the corpus

Word / gloss	Homographs (transliteration and frequency)*
 Mongolia	*mun̄g̅g1ul* 3496535 *mun̄g̅gul* 388308 **mon̄g̅gol** 171610 *mon̄g̅gul* 71666 *mon̄g̅g1ol* 5714 *munggul* 4519 *mun̄g̅gol* 2107 *mon̄g̅g1ul* 383 *mon̄g̅3gol* 224 *monggul* 103 *monggol* 7
 game	**nagadum** 1781301 *nagadom* 19666 *nagaduem* 4068
 song	**daguu** 1258831 *dagou* 86530 *dagoo* 51920 *daguo* 20100

* Italic homographs are *incorrect*. The transliteration scheme used are ours (the specification is available on `http://mandel.cloudtranslation.cc/moncode_en.html`). Number 1-3 in the homographs means free variation selector 1-3 which is used to select a specific presentation form of a Mongolian letter (Unicode 2015).

2 

3 

4 To emphasize the letter *ang* is one code point, we transliterate it as *n̄g̅*. It is often pronounced as [ŋ].

5 http://cloudtranslation.cc/corpus_minority.html

The first entry, *mun̄gg1ul* (meaning **Mongolia**) is the most frequent encoding in the corpus, however, it is spelled wrong (the correct one is *mon̄ggol*)! This is an unfortunate state of affairs, as when shown on the computer screen or printed on the paper, the homographs appear the same, but internally they are different, thus resulting in great headaches for many NLP applications (e.g. for search engines[6]). However, As Andrew West pointed out[7], the Mongolian Unicode standard is here to stay, and we can only live with it now.

We propose two ways to alleviate the problems caused by the current Mongolian Unicode standard [11]: first we developed a Unicode input method for Mongolian which shows the transliteration and thus helps users to select the correct homograph; second, for the bulk of existing Unicoded Mongolian, we normalize the corpus using an automatically compiled homograph table and the **vowel harmony principle** [1][6][10]. To test the effectiveness of our normalization method, we do experiments on our publicly accessible search engine[8] which supports Mongolian and the results show significant improvement on recall. Experiments on statistical Mongolian-Chinese machine translation also result in an absolute improvement of 5.70 BLEU score [9].

2 Yunmeng: A new Unicode Mongolian Input Method

The most widely used Mongolian input methods in China is Menksoft Mongolian IME [7] but it is not based on Unicode (this fact alone shows that Mongolian Unicode Encoding is problematic). There are researches on Unicode Mongolian Input method, e.g. [5][8], but few free Unicoded Mongolian Input software have emerged from the researches.

We have implemented *Yunmeng*, a Mongolian Input based on the Unicode standard. We designed a transliteration scheme compatible with Unicode code standard and the specification is available online. Control characters such as three Mongolian Free Variation Selectors are supported. There are 3 salient characteristics of the Yunmeng input method:

1. To input Mongolian, one only needs to remember the pronunciations of the Mongolian vowels and consonants. No keyboard layout is required.
2. Both letter-by-letter and word-by-word input modes are supported. To input a Mongolian word, just type its shortened transliteration which consists of its first syllable and all the consonants of the remaining syllables, e.g. the word *monggol*'s shortened transliteration is *monggl*. In fact, as soon as its prefix *mongg* is typed, the first candidate is shown correctly. See Fig. 1 for details.
3. The transliteration can be shown along with the traditional Mongolian script. This helps the user to choose the correct encodings. And if an encoding fails to satisfy the vowel harmony (it basically says that all the vowels in a word are either of masculine or feminine gender. The neuter vowel *i* is compatible with both the masculine and

---

6 Prof. Garudi of Inner Mongolia Normal University, personal communication.

7 Personal communication.

8 http://search.cloudtranslation.cc/

the feminine gender), it's highlighted to the user as a possible incorrect encoding. Again refer to Fig. 1 for examples.

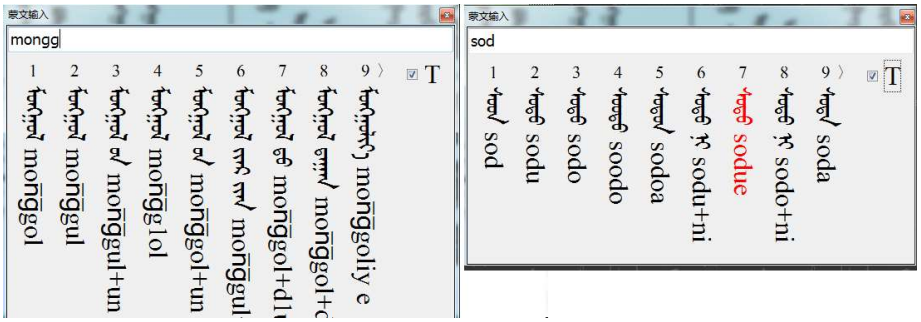


Fig. 1. Yunmeng input interface for words *monggol* and *sodo*.

The Yunmeng input software is freely available for downloading[9]. This input method can help to reduce the incorrect homographs caused by the current Unicode Mongolian encoding.

3 Normalization of the Mongolian Corpus in Unicode

As can be seen from Table 1, there are lots of wrongly encoded Mongolian words in the corpus. This makes the vocabulary unnecessarily large and severe data sparseness will result. If no pre-processing is done, basic NLP processing tasks such as tagging, parsing, machine translation, searching, etc., will suffer a lot in accuracy. The obvious answer is of course **normalization**, that is, to convert a wrongly spelled Mongolian word to its correct normal form.

Now the problem is, how can we find all the (correct or wrong) homographs of a word? It's out of the question to compile the homograph table (like Table 1) manually. There are simply too many variant spellings/encodings which appear the same.

The second problem is, suppose we have gotten such a table, how can we determine which homograph is the correct one? Frequency information cannot be used reliably here as we have seen from Table 1 that the most used form of the word *monggol* is the wrong one!

Our solutions are outlined below:

- For problem 1, we can resort to glyph similarity comparison algorithms to find the homographs that look the same. See [4] for a review for graph matching algorithms.
- For problem 2, as long as the application is not concerned with the morphology of the Mongolian words, it actually does not matter whether a correct homograph can be chosen or not. We can simply pick the most frequent (albeit perhaps wrong) one as the normalized encoding, although word encodings violating the vowel harmony should certainly be discarded. For example, word-based or phrase-based statistical

[9] http://uread.superfection.com/software/uread_mongolian.rar

machine translation normally does not need morphological analysis. However, if morphological information is required, then manual work is unavoidable.

In our current implementation of the homograph mining algorithm, we require that their glyphs be exactly the same if two word encodings are regarded as homographs. This is an unnecessarily *strict* requirement because it will find fewer homographs than our naked eyes can recognize, but it serves our purpose here. The algorithm is outlined below.

Algorithm 1. Homograph mining.

```
foreach word in Vocabulary
  foreach generated homograph candidate according to the
letter equivalence table[10]
    if glyph_same(word, candidate)[11]
    then {
      put the pair into a equivalent group
      if one of the pair already has equivalent groups
      then merge the generated group with others
    }
```

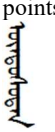
We find 84611 homograph equivalent groups for our 150-million-word corpus! To appreciate this result, here we just give an example of these groups (Yunmeng transliterations, of the Monglian word shown at the footnote[12]):

```
oendosoeden uentusutan uentusueten uentuesuten uentuesuetan uentuesueten
uentosotan uendusutan uendusuten uendusudan uendusuden uendusuetan
uendusueten uendusueden uendoesoden uendoesoeten uenduesodan uenduesoden
uenduesuten uenduesudan uenduesuden uenduesuetan uenduesueten
uenduesuedan uenduesueden uendosotan uendosodan uendosoden uendosueten
uendosueden oentoesoeten oentosotan oendusudan oendusuden oendusoeden
oendoesoeten oendoesoedan oendoesoeden oenduesueten oenduesoeten
oendosoden uend1uesueten uend1uesud1an oenduesued1en uend1usud1an
uend1uesued1en
```

For automatic normalization for use in search engines and statistical machine translation, we simply choose the most frequent variation as the normalized encoding (upon applying vowel harmony), if our small normalization table compiled manually does not

[10] The letters *o* and *u* are regarded *equivalent* letters. We collected 22 such equivalent pairs and they form the letter equivalence table. Note that equivalent letters do not always have same glyphs in all positions, e.g. some letters are only equivalent at the medial positions.

[11] We rely on the Microsoft Uniscribe engine to generate the correct glyphs. However, as [1] points out, even if Microsoft failed to generate some of the correct glyphs.

[12] 

show which encoding is the correct one. We report our experiments of the normalization in the next section.

4 Efficacy of the Normalization

To test the effectiveness of the normalization, we do experiments both on our search engine which supports Mongolian, and on statistical Mongolian-Chinese machine translation:

- For the search engine, we do normalization on both the query and documents.
- For the machine translation, we also do normalization on both training and testing data.

The result of the search engine experiment is reported in Table 2. We randomly choose a few words with various frequencies and search these words in the Google search engine and ours[13]. Because there are many Mongolian sites still not indexed by our search engine[14], for meaningful comparison, we restrict Google search results using the "site:" directive.

Table 2. Effects of Mongolian normalization on search engine results

Word and frequency	Google hits	Our search engine hits before normalization	Our search engine hits after normalization
(2)	18	6	**16128**
(15)	1	1	**1**
(33)	17	28	**283**
(166)	125	78	**209**

It can be seen that the recall of the search engine is boosted significantly. It is safe to say that we have made a big step toward solving the search problem that has been plaguing the Unicoded Mongolian for many years!

The result on statistical Mongolian-Chinese machine translation is shown in Table 3. The system used is an in-house phrase-based machine translation system. The training data are the China laws and government reports, with 59k parallel sentences, and only one reference translation is provided for test data. It can be seen that an absolute improvement of 5.70 BLEU-4 score can be achieved on average upon normalization. The new Mongolian-Chinese machine translation system with normalization is also available online[15].

[13] Google's powerful engine can index pdf files, while ours does not yet.

[14] For this experiment, we only indexed two popular Mongolian website: www.mgyxw.net and mgl.nmg.gov.cn

[15] http://cloudtranslation.cc/mt

Table 3. Effects of Mongolian normalization on machine translation

	Size (sentences)	BLEU before normalization	BLEU after normalization
Training data	59K		
Test data1 (government report)	266	38.44	**46.07**
Test data2 (leader's speech)	122	50.28	**56.54**
Test data3 (law)	123	57.19	**60.41**
Average improvement			+5.70

5 Conclusion

We discussed the *homograph plague* problem that has defied the widespread use and application of Unicoded traditional Mongolian for many years, and we propose a new input method called **Yunmeng** and a normalization algorithm to deal with it. Experiments on search engines and statistical machine translation showed great improvements.

Acknowledgements. The work done in this paper is partially supported by the Research Fund for the Doctoral Program of Higher Education of China (No. 20130121110040), National High-Tech R&D Program of China (No. 2012BAH14F03), and the Special Fund Project of Ministry of Education of China (Intelligent Conversion System from Simplified to Traditional Chinese Characters). We thank Dr. Yanlong He for kindly providing the Mongolian-Chinese test corpus for the statistical machine translation experiment.

References

1. Batjargal, B., Khaltarkhuu, G., Kimura, F., Maeda, A.: A Study of Traditional Mongolian Script Encodings and Rendering: Use of Unicode in OpenType Fonts. Int. J. of Asian Lang. Proc., 21(1), 23-44 (2011)
2. Chinggaltai: A Grammar of the Mongol Language. Frederick Ungar Publishing Co, New York (1963)
3. Choijinzhab: Mongolian Encoding. Inner Mongolia University Press, Hohhot. (确精扎布：蒙古文编码．内蒙古大学出版社，呼和浩特) (2000) (in Chinese) http://www.babelstone.co.uk/Mongolian/MGWBM.html
4. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty Years of Graph Matching in Pattern Recognition. International Journal of Pattern Recognition and Artificial Intelligence, 18(03), 265-298 (2004)
5. Daoerji, F., Fengshan, B., Huijuan, W. U.: Research on Mongolian Input Method in Unicode. Journal of Chinese Information Processing, 24(6), 120-124+128 (2010) (in Chinese)
6. Goldsmith, J.: Vowel Harmony in Khalkha Mongolian, Yaka, Finnish and Hungarian. Phonology, 2(01), 253-275 (1985)

7. MünggeGal: Menksoft Mongolian IME, http://www.menksoft.com/
8. Ochir, Wang, G. F.: Corpus and Mongolian Inputting Methods. In: International Conference on Chinese Computing 2005, Singapore (2005)
9. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318. Association for Computational Linguistics (2002)
10. Poppe, N.: Grammar of Written Mongolian. Otto Harrassowitz Verlag, Wiesbaden (1974)
11. The Unicode Consortium: The Unicode Standard, Version 8.0.0, (2015) http://www.unicode.org/versions/Unicode8.0.0/